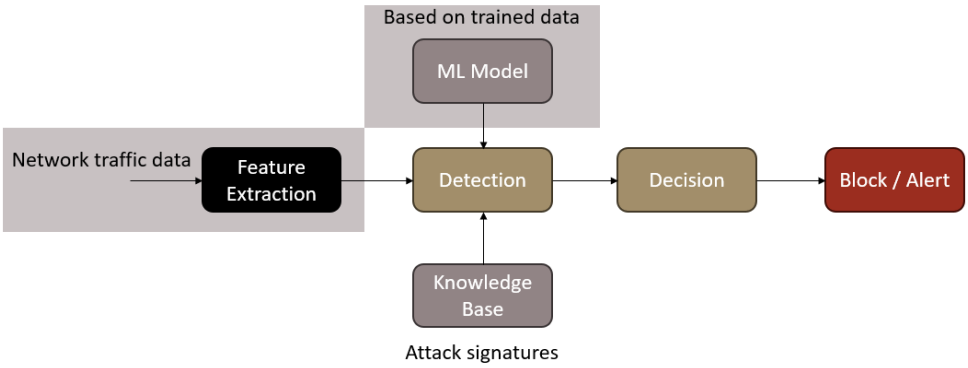
Report on the Jupyter Notebook: Initial Experiment and Evaluation Setup

1. Objective

The notebook demonstrates a data preparation and machine learning pipeline for classifying network traffic as either Benign or DDoS (Distributed Denial of Service). It includes dataset sampling, preprocessing, feature selection, correlation analysis, and model training using a Multilayer Perceptron (MLP) neural network.



2. Data Loading and Sampling

The dataset was loaded from Google Drive and processed in chunks of 100,000 rows, sampling 10% from each chunk. The final dataset contains 1,279,463 entries and 85 columns. The class distribution is approximately balanced between DDoS (647,297 samples) and Benign (632,166 samples).

3. Feature Overview and Reduction

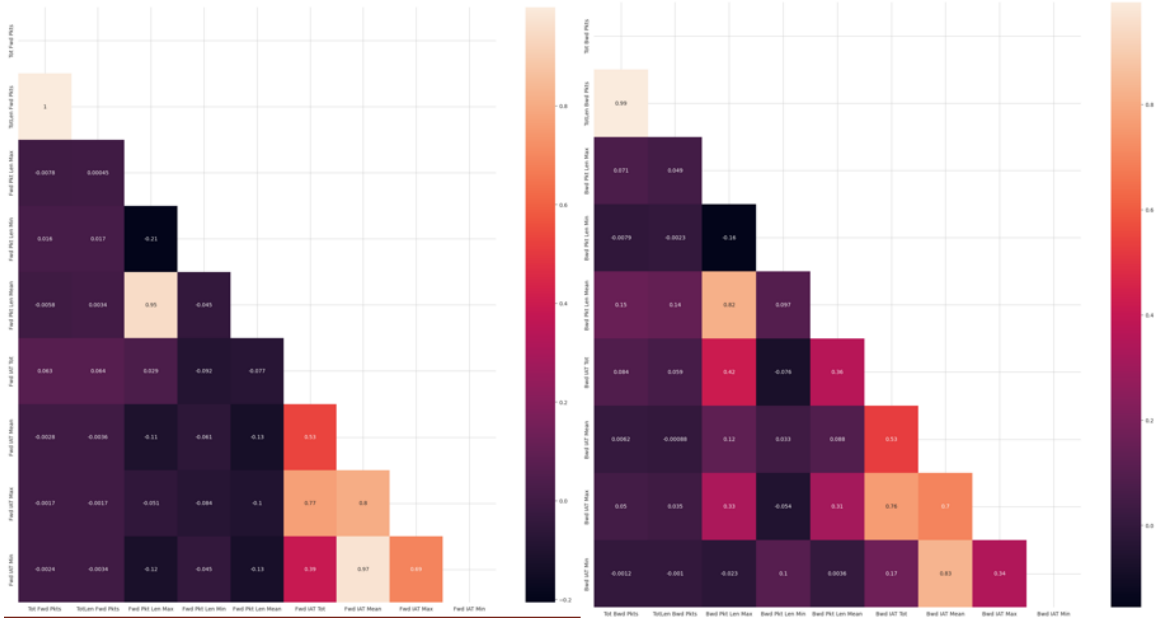
Each record describes a network flow with attributes such as flow identifiers, packet statistics, timing metrics, and TCP flag counts. Due to hardware constraints, the notebook removes redundant or less important features in stages:
- Level 1: Ambiguous or redundant (e.g., Pkt Len Std, Pkt Len Var)
- Level 2: High-cost computations (e.g., IAT Std, Active/Idle stats)
- Level 3: Derived metrics (ratios, averages, rates)
- Level 4: Transport-specific metrics (subflow or segment size)
- Level 5: Irrelevant or low-utility signals (flag counters, averages)
After filtering, 36 columns remain, including 6 identifiers and 29 feature columns.

4. Data Preparation and Cleaning

The reduced dataset is stored in df_filtered. Non-numeric identifiers are removed, missing values dropped, and labels encoded (0 for Benign, 1 for DDoS). The cleaned dataset has 26 columns, all numeric.

5. Feature Correlation Analysis

Correlation heatmaps identify highly correlated features such as TotLen Fwd Pkts, TotLen Bwd Pkts, Fwd Pkt Len Mean, and Fwd IAT Min, which are then removed. This ensures reduced redundancy and multicollinearity, leaving 32 columns.



6. Data Transformation

The dataset is converted to NumPy arrays: X (features) and Y (one-hot encoded labels). Data is standardized and split into training (75%) and testing (25%) sets, totaling 959,597 and 319,866 samples respectively.

```
data_clean.head()
```

	Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	Fwd Pkt Len Max	Fwd Pkt Len Min	Bwd Pkt Len Max	Bwd Pkt Len Min	Bwd Pkt Len Mean	Flow IAT Mean	Flow IAT Max	...	Bwd IAT Max	Bwd IAT Min	Pkt Len Min	Pkt Len Max	Pkt Len Mean	FIN Flag Cnt	SYN Flag Cnt	RST Flag Cnt	ACK Flag Cnt	Label
0	5032233	3	5	603.0	0.0	972.0	0.0	194.400000	7.188904e+05	5004303.0	...	5004614.0	6.0	0.0	972.0	175.000000	0	1	0	0	1
1	10993127	3	5	408.0	0.0	972.0	0.0	194.400000	1.570447e+06	5990709.0	...	5990709.0	3.0	0.0	972.0	153.333333	0	1	0	0	1
2	479425	3	7	142.0	0.0	1460.0	0.0	671.714286	5.326944e+04	232452.0	...	232857.0	227.0	0.0	1460.0	440.363636	0	1	0	0	1
3	9042123	1	1	0.0	0.0	0.0	0.0	0.000000	9.042123e+06	9042123.0	...	0.0	0.0	0.0	0.0	0.000000	0	0	0	1	1
4	1907579	1	1	0.0	0.0	0.0	0.0	0.000000	1.907579e+06	1907579.0	...	0.0	0.0	0.0	0.0	0.000000	0	0	0	1	1

5 rows × 26 columns

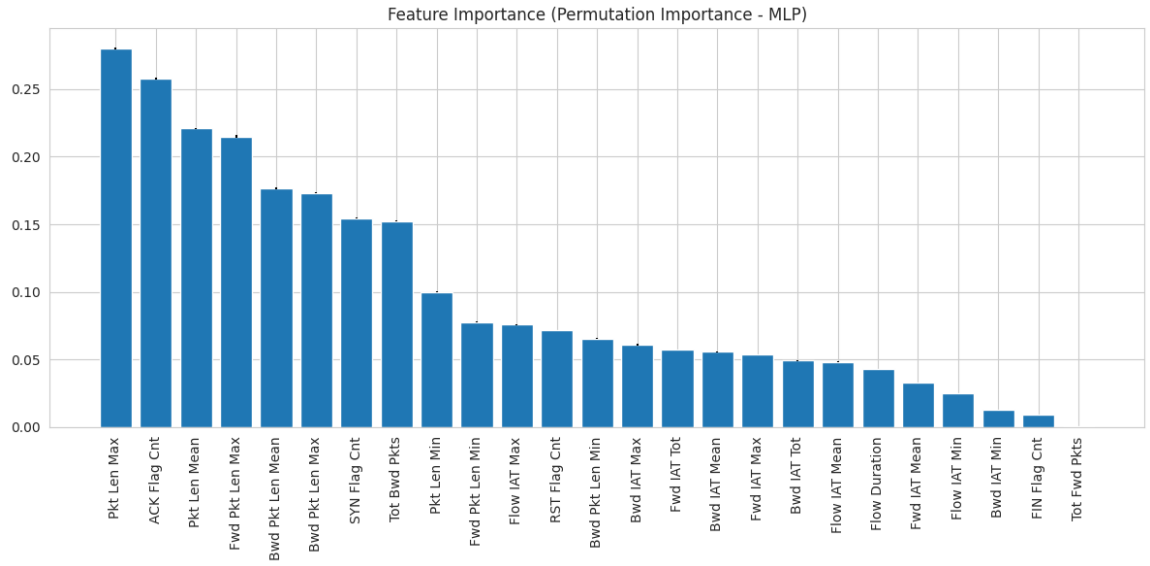
7. Model Training

A Multi-Layer Perceptron (MLPClassifier) with two hidden layers (64 and 32 neurons) is trained using ReLU activation and the Adam optimizer for 200 epochs. The model achieves an accuracy of 97.54% with balanced precision and recall:
- Precision: 1.00 (Benign), 0.96 (DDoS)
- Recall: 0.95 (Benign), 1.00 (DDoS)
- F1-score: 0.97 (Benign), 0.98 (DDoS)

```
Accuracy: 0.9754365890716737

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.95      0.97    157931
           1       0.96      1.00      0.98    161935

   micro avg       0.98      0.98      0.98    319866
   macro avg       0.98      0.98      0.98    319866
weighted avg       0.98      0.98      0.98    319866
 samples avg       0.98      0.98      0.98    319866
```

8. Feature Importance

Permutation importance shows that features such as Pkt Len Max, ACK Flag Cnt, Fwd Pkt Len Max, Bwd Pkt Len Max, and SYN Flag Cnt contribute most to model accuracy, indicating packet size and flag behavior are strong indicators of DDoS traffic.



9. Key Insights

- Sampling and chunking efficiently manage large datasets.
- Feature reduction minimizes computation without sacrificing performance.
- The MLP achieves 97.5% accuracy with robust generalization.
- Packet length and TCP flag metrics are key indicators of DDoS activity.

10. Conclusion

The notebook successfully implements a complete network intrusion detection pipeline. It efficiently processes large-scale data, selects meaningful features, and trains a high-performing neural classifier. Future work could explore multi-class attack detection, deeper neural models, or real-time implementation.